Analyzing cancer forum discussions with text mining

Gerard van Oortmerssen[1], Stephan Raaijmakers[2], Maya Sappelli[2], Erik Boertjes[2], Suzan Verberne[3,4], Nicole Walasek[3], Wessel Kraaij[2,4]

[1] Tilburg University
[2] TNO, The Hague
[3] Radboud University, Nijmegen
[4] LIACS, Leiden University

Abstract. We present a multilingual, open source system for cancer forum thread analysis, equipped with a biomedical entity tagger and a module for textual summarization. This system allows users to investigate textual co-occurrences of biomedical entities in forum posts, and to browse through summaries of long discussions. It is applied to a number of online cancer patient fora, including a gastro-intestinal cancer forum and a breast cancer forum. We propose that the system can serve as an extra source of information for medical hypothesis formulation, and as a facility for boosting patient empowerment.

1 Introduction

Online patient communities are a potentially valuable source of information for cancer patients. In these communities, patients share detailed information on their disease, treatment, side effects of treatments and coping strategies, as well as their experienced quality of life. The aggregated information from the entire history of discussions can contribute to patient empowerment, but may also inspire clinical hypotheses. This short paper[5] presents an open source system[6] for the automated analysis of cancer forum posts supported by text mining.

2 System architecture

Our system has three main components: a pipeline for analyzing entities and relations in forum posts, a summarization module for summarizing discussion threads, and a user interface for explorative search. This section describes the three components in detail. The system currently contains data from three forum communities: the Dutch Breast Cancer community (BVN), the Facebook community *GIST Support International*, and the medical section from the Dutch Viva forum[7]. In order to display potential relations between elements such as

[5] This work is supported by an SIDN Fund grant (`https://www.sidnfonds.nl`).

[6] `https://github.com/patientforumminer/PFM`.

[7] `https://borstkanker.nl`, `http://www.gistsupport.org`, `http://forum.viva.nl`

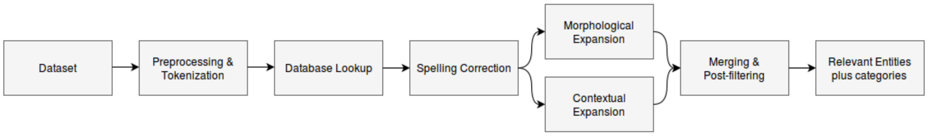
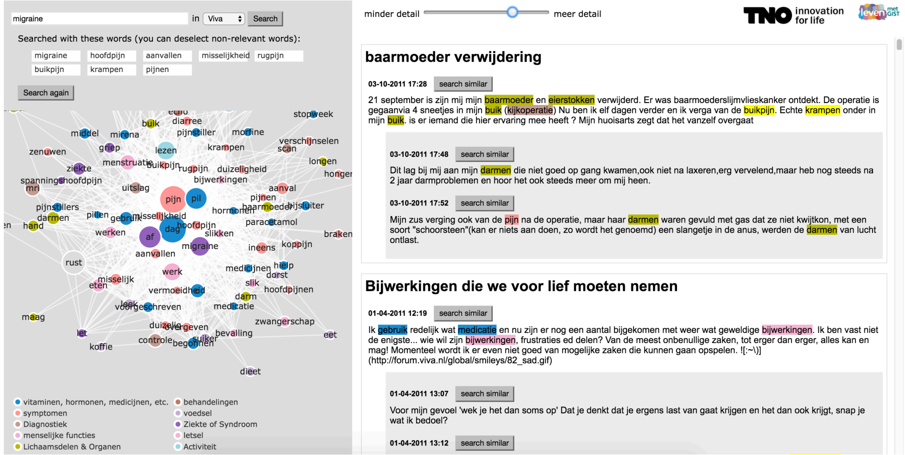


Fig. 1. Overview of the individual steps of the entity tagging pipeline.

side effects and treatments, textual entities (names and concepts) are tagged with their respective medical (semantic) categories, for both Dutch and English. As a first step, the system preprocesses forum post threads. Data is lower-cased, URLs and all non-alphanumeric characters are removed except for hyphens and commas. Weights and dosages are extracted from the threads based on regular expressions and directly tagged as such. After this step, all numbers are removed, followed by tokenization and stop word removal. Subsequently, a database lookup is executed for all single terms, using the Unified Medical Language System (UMLS) which is in English. Each Dutch term is first translated into English using a translation dictionary extracted from DBpedia[8]. For every input term, the semantic types are extracted from UMLS and the most frequent one is chosen as the category for the term. If a particular term cannot be matched within the UMLS database, the DBpedia database is queried for that term and the most specific type is extracted. If this lookup also fails, the Medical Subject Headings (MeSH) database is queried and the term is matched to the first topical descriptor for an exact string match with the broader descriptor. The next step is to apply spelling correction to the remaining terms that have not yet been matched in the first look-up step. Only unmatched terms with a low frequency in the corpus (≤ 2) are considered. These either represent true misspellings or rare morphological variants of the unmatched types. For each of these low-frequency terms, the matched entity with the lowest weighted relative edit distance is determined. Character changes at the beginning of the term are prohibited. To increase the number of matched entities we include morphological variances based on lemma matching (using *Pattern* [2]) for terms with more than 4 characters. Moreover we expand the number of entities using contextual relations. For this purpose, a Word2Vec model [1] was trained on all data[9]. The category of the majority of the 5 closest neighbors was assigned to each unmatched term, if a threshold of 3 was exceeded. Since lemma lookup and Word2Vec expansion are executed simultaneously, the results were merged. In case of disagreement, the result of Word2vec was preferred. As a final step, a selection of categories for the application was made. Moreover, high-frequency terms (such as ‘sleep’ or ‘live’) were excluded from the results, since tagging these terms was not relevant for the application. We evaluated the system in terms of the precision for the identi-

[8] Available from `http://downloads.dbpedia.org/2015-10/core/`.

[9] Parameter settings: model=CBOW, feature dimensionality= 500, window size=3, minimum word count=3, number of cores=3, other parameters are set to default values.



migraine in Viva

Searched with these words (you can deselect non-relevant words):

migraine hoofdpijn aanvallen misselijkheid rugpijn
buikpijn krampen pijnen

minder detail meer detail

baarmoeder verwijdering

03-10-2011 17:28

21 september is zijn mij mijn baarmoeder en eierstokken verwijderd. Er was baarmoederslijmvlieskanker ontdekt. De operatie is gegaanvia 4 sneetjes in mijn buik (kijkoperatie) Nu ben ik elf dagen verder en ik verga van de buikpijn. Echte krampen onder in mijn buik. is er iemand die hier ervaring mee heeft ? Mijn huoisarts zegt dat het vanzelf overgaat

03-10-2011 17:48

Dit lag bij mij aan mijn darmen die niet goed op gang kwamen,ook niet na laxeren,erg vervelend,maar heb nog steeds na 2 jaar darmproblemen en hoor het ook steeds meer om mij heen.

03-10-2011 17:52

Mijn zus verging ook van de pijn na de operatie, maar haar darmen waren gevuld met gas dat ze niet kwijtkon, met een soort "schoorsteen"(kan er niets aan doen, zo wordt het genoemd) een slangetje in de anus, werden de darmen van lucht ontlast.

Bijwerkingen die we voor lief moeten nemen

01-04-2011 12:19

Ik gebruik redelijk wat medicatie en nu zijn er nog een aantal bijgekomen met weer wat geweldige bijwerkingen. Ik ben vast niet de enigste... wie wil zijn bijwerkingen, frustraties ed delen? Van de meest onbenullige zaken, tot erger dan erger, alles kan en mag! Momenteel wordt ik er even niet goed van mogelijke zaken die kunnen gaan opspelen. ![:~\)](http://forum.viva.nl/global/smileys/82_sad.gif)

01-04-2011 13:07

Voor mijn gevoel 'wek je het dan soms op' Dat je denkt dat je ergens last van gaat krijgen en het dan ook krijgt, snap je wat ik bedoel?

01-04-2011 13:12

Fig. 2. A screenshot of the web-based interface with in the top-left corner the query box, below it the entity graph, and on the right side the results that are retrieved for the query (in Dutch).

fication and classification of the most frequent 300 matched entities. For the first task, the average precision across annotators was 0.79 (inter-rater agreement in terms of Cohen's $\kappa = 0.77$). The classification task yielded an average weighted precision per category of 0.74 ($\kappa = 0.67$).

The search module of the system returns posts in the context of a discussion thread, often consisting of dozens or even hundreds of posts. We automatically summarize the discussion threads with *extractive summarization*: showing the most relevant sentences in the thread while hiding the less relevant sentences in between them. Input for this summarization is a ranking of the sentences by their relevance for a summary. For the prediction of relevance, we trained a linear regression model on human[10] reference summaries created for the English and Dutch forum data. In the model, we used the number of raters that selected a sentence as outcome variable (a sentence selected by 4 or 5 raters is more relevant than a sentence selected by 1 or 2 raters). As independent variables we used a number of generic sentence features such as the position in the thread, the sentence length and the similarity with the full thread. We performed a blind side-by-side comparison of the model's summaries with human-created summaries, which showed that our model's summary was judged as equally good as or better than the human-created summary [4].

The graphical user interface of the system allows for an iterative search process in which the user quickly reaches relevant search results, supported by query expansion, entity tagging and automatic thread summarization. Figure 2 shows the system's GUI. It is divided into two main parts, the left part supporting

[10] We used 7 raters: 5 non-experts and two experts. More information about the reference summaries and the summarization module was published in [3].

the querying process, the right part for browsing search results. The user typically starts with entering one or more keywords (upper left), that are expanded with related terms by the system using Word2Vec trained on all patient forum data. The system presents the expansion terms to the user, allowing to decide whether or not to include these terms in a next query. The term network shows terms that occur in the search results and their inter-connection. Two terms are connected when they co-occur frequently in the same context (e.g. message) in the result set. The strength of the relationship is depicted by the thickness of the edge in the network graph. Each node is coloured according to its classification (e.g. medicine, food, symptom), and a node's size is proportional to the number of occurrences in the results. The network facilitates the discovery of unexpected links between terms (e.g. a food substance mentioned frequently in combination with a symptom). This is typically useful for expert users looking for new medical hypotheses. The right hand side of the GUI shows the search results. These are threads, consisting of the first message in a thread, followed by a list of comments on that message. The opening post of the thread is always shown; sentences of other posts in the thread are only shown if the user prefers to see more detail, governed by the slider on top of the screen. Which sentences are shown first when the slider is moved to the right is decided by the summarization module.

3 Conclusions

We have presented an open source system for the automated analysis and interactive inspection of cancer forum posts. For the purpose of knowledge extraction from patient-generated forum data, our future work will focus on techniques for matching new forum threads with existing ones, and on connecting user-generated content to moderated content, such as curated taxonomies and published medical information. Further, we will address query relevance for summarization, and the thorough evaluation of our system in task-based settings.

References

1. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
2. Smedt, T.D., Daelemans, W.: Pattern for Python. Journal of Machine Learning Research 13(June), 2063–2067 (2012)
3. Verberne, S., van den Bosch, A., Wubben, S., Krahmer, E.: Automatic summarization of domain-specific forum threads: collecting reference data. In: Proceedings of The ACM SIGIR Conference on Human Information Interaction & Retrieval (CHIIR). pp. 253–256 (2017)
4. Verberne, S., Krahmer, E., Hendrickx, I., Wubben, S., van den Bosch, A.: Creating a Reference Data Set for the Summarization of Discussion Forum Threads. Language Resources and Evaluation pp. 1–23 (2017)